Exploring Fact-Focused Relevance and Novelty Detection

Jahna Otterbacher
Department of Public and Business Administration
University of Cyprus
Nicosia, Cyprus

Dragomir Radev
School of Information, Departments of EECS and Linguistics
University of Michigan
Ann Arbor, Michigan, USA

Exploring Fact-Focused Relevance and Novelty Detection

Abstract

Purpose – Automated sentence-level relevance and novelty detection would be of direct benefit to many information retrieval systems. However, the low level of agreement between human judges performing the task is an issue of concern. In previous approaches, annotators were asked to identify sentences in a document set that are relevant to a given topic, and then to eliminate sentences that do not provide novel information. Currently, a new approach is explored in which relevance and novelty judgments are made within the context of specific, factual information needs, rather than with respect to a broad topic.

Design/methodology/approach – An experiment is conducted in which annotators perform the novelty detection task in both the topic-focused and fact-focused settings.

Findings – Higher levels of agreement between judges are found on the task of identifying relevant sentences in the fact-focused approach. However, the new approach does not improve agreement on novelty judgments.

Originality/value – The analysis confirms the intuition that making sentence-level relevance judgments is likely to be the more difficult of the two tasks in the novelty detection framework.

Keywords Relevance judgments, Novelty, Human factors

Paper type Research paper

Introduction

A core challenge for future information retrieval (IR) systems is to find information that is not only relevant to a user's need, but is also novel (Allan, 2005). To this end, “novelty detection,” the task of identifying units in a text or set of texts that express interesting and previously unseen information, has been introduced. In contrast to systems that retrieve all relevant items given a user's topic of interest, systems incorporating novelty detection aim to reduce the amount of redundant information presented to the user.

Several recent research problems focus on finding novelty at fine levels of textual granularity and in particular, at the sentence level. For instance, Allan and colleagues proposed -summarization, in which summaries of a set of incoming documents are produced over time (Allan, 2001). More specifically, -summaries highlight what has changed since the previous point in time. Their objective is to find “interesting” sentences, which are both useful, or relevant to the desired topic, and novel. The sentences identified as being interesting are then used in producing extractive summaries.

Clearly, a means for automatically detecting relevant, novel information at the sentence level would be of direct benefit to extractive text summarizers, as controlling the amount of redundancy while still choosing the most relevant sentences is a well-known problem in summarization research (Goldstein, 1999). In addition, novelty detection would be useful in the context of question-answering systems that, after having identified documents relevant to the user's input question, then find relevant sentences and perform answer extraction from the selected sentences. Such systems might employ novelty detection techniques in order to determine which sentences express the same answer to a given question and which sentences provide new or unique answers.

TREC: sentence-level novelty detection

A major initiative towards the development of sentence-level novelty detection systems was the Text Retrieval Conference (TREC) Novelty Track, conducted as a part of the 2002, 2003

and 2004 programs[1]. Its goal was to train systems that perform a two-stage task. Given a topic query and a set of relevant documents, the systems should first retrieve all sentences that are relevant to the topic. In the second step, the system should choose, from the list of relevant sentences, the novel sentences, defined as those containing "previously unseen information" (Soboroff, 2003).

Several challenges were noted by the organizers with respect to creating manually-labeled data sets for training and evaluating systems. Most notably, in 2002, the judges choose very few relevant sentences, which resulted in many negative and few positive relevance examples (Harman, 2002). At the same time, most relevant sentences were also novel. The organizers cited document relevance as an issue, since they presented a set of pre-selected documents to the assessors. Therefore, in 2003, several improvements were made (Soboroff, 2003). One assessor, who was deemed the official judge, created the topics and identified a set of 25 relevant articles by searching a collection of news documents. The documents were then ordered chronologically, and the assessor performed the two-stage retrieval process. A second assessor also performed the task in order to assess the level of agreement. This time, the distributions of sentences marked as being relevant and novel were more reasonable. However, a large assessor effect was noted. The level of agreement on relevance judgments varied across topics, and in particular, tended to be lower for topics describing opinions rather than events (Soboroff, 2005).

Over all 50 topics in the 2003 data, the median proportion of sentences marked as being relevant by the official assessor that were also chosen by the second assessor was 0.69. The median agreement by the second judge on novelty judgments was 0.56. In 2004, the proportion of agreement on relevance and novelty judgments over the 50 topics used was slightly less, at 0.60 and 0.35, respectively. For the 2004 relevance judgments, Soboroff and Harman (2005) reported a corresponding Kappa statistic of 0.549. Kappa corrects the level of agreement for random occurrences, such that a Kappa of zero indicates that agreement is no better than chance (Carletta, 1996) (Cohen, 1960). The 2004 result indicates that the level of agreement between the two judges on relevance judgments is significantly greater than what

[1] http://trec.nist.gov

would be expected by chance. However, as noted by the organizers, according to most scales of interpretation for Kappa, this corresponds to a low-to-moderate level of agreement between judges (Krippendorff, 1980) (Landis, 1977).

In an attempt to improve on the TREC results, Schiffman (2005) built a corpus of novelty judgments using a modified procedure. Assessors did not make relevance judgments, and were shown only two news articles at a time. One was presented as background information, while the other was considered a new document. The judges then chose the spans of text in the latter document that presented novel information not contained in the background article. The new task yielded rather low rates of agreement between assessors (Kappa of 0.24). This result illustrates the difficulty of the novelty task, even when the problem is constrained such that a small number of related documents are considered.

Variations in relevance judgments

As reported by the TREC organizers (Soboroff, 2005), the results on the manual novelty tasks are not surprising, given that relevance judgments are known to vary significantly across assessors (Voorhees, 1998). In addition, it has been noted that the level of agreement on identifying novel information is dependent on that of the relevance task, and that the relevance task is likely the more difficult of the two (Allan, 2003). For example, it has been shown that while users can typically identify highly relevant items rather easily, they often struggle when making relevance assessments of marginally relevant and less relevant items (Vakkari, 2004).

Despite that the concept of relevance is essential for the development and evaluation of IR systems, its nature is still not well understood, nor is it always clear how to operationalize relevance within a given system (Froehlich, 1994) (Mizzaro, 1997). It is also well known that human judgments of relevance vary, both across multiple judges and over time by the same judge (Schamber, 1994), leading some to criticize the use of such judgments (e.g. (Cuadra, 1967) (Harter, 1996)).

In evaluating system performance, several studies have suggested that the variation across assessors does not significantly alter the resulting system rankings. For example, in an experiment using the TREC-4 data, Voorhees found that the resulting system rankings produced using relevance judgments collected from different assessors were highly correlated (Voorhees, 1998). Therefore, from a systems evaluation perspective, how well assessors agree on relevance judgments may not cause too much of a concern.

However, from a system building standpoint, low interjudge agreement is a problem. In order to be able to train systems that replicate human judgment on a given task, one must first verify that humans themselves produce similar judgments (Carletta, 1996). In fact, in evaluating automated approaches to a classification task, the agreement between independent human judges on the task typically represents an upper bound for system performance (e.g. (Marcu, 2002) (Teufel, 2002)). Therefore, in order to make progress in developing systems for sentence-level relevance and novelty detection, it is desirable to start with manually annotated data on which a satisfactory level of interjudge agreement has been established.

Goals of the current work

Currently, we investigate a new approach to sentence-level relevance and novelty detection that is fact-focused. In contrast to the topic-focused case, we propose to constrain the task such that judgments are made in the context of specific facts surrounding the topic. We should note that while the TREC evaluation considered sets of news articles describing both events and opinions, our work focuses on the former only. While it may also be possible to represent an opinion topic as a set of specific questions, this approach is beyond the scope of our current work.

Our first goal is to examine the fact-focused approach to the novelty task, as to whether it results in satisfactory levels of interjudge agreement, or reproducibility. To this end, we

present an experiment in which we asked annotators to complete both the fact-focused and topic-focused novelty tasks. In addition, we will consider the differences and similarities between the fact-focused and topic-focused approaches, by comparing the sets of sentences selected under the two settings.

Fact-focused novelty detection

In the fact-focused scenario, the user has a topic of interest, and has identified a set of relevant documents. Next, we assume that the user has a set of facts of specific interest about the topic. For simplicity, each fact may be stated as a question. Given a fact of interest, the first step is to identify the sentences in the documents that contain relevant information. A sentence contains relevant information only if it provides an answer to the question. In the second step, only the sentences containing previously unseen information about the factual question are kept, thus creating a set of novel sentences. Another way of stating this approach is that a topic is described by a set of factual queries, which represents the most salient information about the event described in the set of news articles. Rather than asking assessors to consider multiple facets of a topic simultaneously, we ask them to focus on specific facts individually, when making relevance and novelty assessments.

We are interested in evaluating this new approach for several reasons. First, previous research has evaluated the agreement between annotators on identifying fact-like semantic units in text with some promising results. For example, in a study by van Halteren and Teufel (2003), independent judges identified factoids contained in 50 texts. The factoids were defined as being “atomic information units" that were represented in First Order Predicate Logic-style semantic expressions. Of relevance to our work is the fact that a high level of agreement between the two judges was achieved, despite that “very short guidelines” were established for how to identify factoids. On the task, both precision and recall were reported to be 96%.

Similarly, in work by Nenkova and Passonneau (2004), assessors labeled Summarization Content Units (SCUs) contained in a given text. The SCUs are fine-grained, clause-like

semantic units. The researchers noted a high level of agreement on annotating the SCUs present in a set of texts (a Krippendorff's Alpha of 0.81, where values above 0.67 indicate strong reliability (Krippendorff, 1980)). The results of such previous studies are promising evidence that annotators agree to a large extent as to which textual units express a particular fact.

Another reason for proposing the fact-focused approach is that clear criteria for labeling relevant sentences can be stated. Since the factual information need is expressed in the form of a question, a relevant sentence should provide an answer. Likewise, a sentence judged as being novel should contain a previously unseen answer to the question. We therefore hypothesize that the reliability of relevance and novelty judgments should be better in the fact-centered approach as compared to the case in which judges seek sentences in the context of a more general topic. It has also been previously noted that the concept of novelty is very context-dependent (Allan, 1999). Therefore, we want to evaluate the case in which judgments are made with respect to a single fact, as compared to a topic, which can be viewed as a set of many facts.

Experimental setup and hypotheses

The annotation experiment was designed to test the following hypotheses:

H1: Annotators will achieve higher levels of agreement in finding sentences relevant to specific factual questions, as compared to finding sentences relevant to a broad topic.

H2: The judges will achieve higher levels of agreement if they are asked to find novel sentences with respect to a fact, as compared to the topic-focused setting.

In addition, we wish to examine the nature of the two-step novelty detection task by comparing the sentences selected under the two approaches. It may be the case that assessors in the topic-focused setting actually take a fact-focused approach when finding

topic-relevant sentences. In other words, when instructed to find sentences that are relevant to a topic, they might search for sentences describing key facts surrounding the news event, even when not explicitly asked to do so. At the same time, we wished to see the extent to which a set of pre-determined factual questions about a news story can represent the essence of its topic. Therefore, we examined the similarities and differences between the sentence sets selected in the topic-focused versus the fact-focused settings.

Data

The data used in the experiment were event topics N4 and N33 from the 2003 TREC Novelty data (Soboroff, 2003). Topic N4 concerns an Egypt Air plane crash while N33 details the sinking of a Russian submarine. The reason for choosing this subject matter is that such stories have a dynamic element, with the facts surrounding them changing over time, such that being able to identify both relevant and novel information over time is important for understanding them.

The authors read through the documents about each story, and came to a consensus on a set of ten questions. The questions concern facts that are central to the stories, that may change with time as news sources publish additional information, and that have atomic answers such as a number, name of a person, or a place. The set of questions created for N4 and N33 are shown in Tables I and II, respectively. In addition, the expected answer types to each question are shown in the square brackets. The corresponding topic queries for the two clusters are shown in Figures 1 and 2[2].

1. How many people were on board? [number]
2. What was the origin of the plane? [place]
3. Where was the flight's destination? [place]
4. What type of aircraft was the plane? [mark/brand]
5. Where did the plane crash? [place]
6. When did the crash occur? [time]

[2] To conserve space, only the title and description portions of the queries are shown, although the narrative portions were given to the judges as well.

7. What was the problem with the plane? [reason/cause]
8. Where was the flight data recorder found? [place]
9. How late was the plane taking off in New York? [time duration]
10. How high was the plane flying? [height]

Table I: Factual questions about the Egypt Air crash (topic N4).

1. How many seamen were on the submarine? [number]
2. How was the submarine damaged? [other]
3. What caused the Kursk submarine to sink? [reason/cause]
4. When did the Kursk submarine sink? [date]
5. Where did the Kursk sink? [place]
6. What time did Americans record the sound of an explosion? [time]
7. How far down did the Kursk sink? [depth]
8. Who is the Russian defense minister? [name]
9. Where was Putin during the rescue operation? [place]
10. Which U.S. submarines were in the Barents Sea when the Kursk sank? [name]

Table II: Factual questions about the sinking of the Kursk submarine (topic N33).

Title: Egyptian Air disaster 990 **Description:** Egyptian Air Flight 990 disaster in October of 1999.

Figure 1: TREC topic query for N4.

Title: Russian submarine Kursk sinks **Description:** The Russian submarine Kursk sank in the Barents Sea killing all 118 aboard in August 2000.

Figure 2: TREC topic query for N33.

Subjects

Six paid subjects participated in the experiment. Three were randomly assigned to the test (fact-based) setting and three to the control (topic-based) setting. As in the TREC annotations, we deemed the first subject in each group to be our official assessors. We will refer to judge ”A" as the official assessor in the control group, and judge “D" as the official judge for the test setting. Judges B and C, and E and F, were used to assess the level of agreement in each of the two settings, respectively.

Each judge performed the same assigned task on the two topics, although in different orders, so that learning effects were controlled. In both settings, judges were given the documents for a given story, which were numbered from 1 to 25 in chronological order. In particular, the

25 documents given to the judges for each story were those deemed as being relevant to the respective topic by the official TREC Novelty assessor[3].

The directions for each group of judges were as follows:

Control Group: Familiarize yourself with the story by reading the topic query and by skimming through the articles. Next, read them carefully in chronological order, recording the document number, sentence number and text of each sentence you find that is relevant to the stated topic. When you are finished finding the set of relevant sentences, make a copy of your data. Now, reread through the sentences that you marked as being relevant, and eliminate those that do not contain novel information. Novel information is "information that has not been previously seen."

Test Group: Read through the set of 10 questions. Familiarize yourself with the story by skimming through the set of articles. Beginning with question one, read through the documents carefully in chronological order, recording the document number, sentence number and text of each sentence you find that provides an answer to the question (a relevant sentence). When you are finished finding relevant sentences, make a copy of your data. Now, reread through the sentences that you marked as being relevant, eliminating those that do not provide novel answers. Novel answers are those that have not been previously seen. Use the same procedure for each of the 10 questions.

Data validation

As previously mentioned, the TREC 2002 data was troubled by the fact that very few sentences were marked as being relevant by the judges, and thus, almost all relevant sentences were also novel. Soboroff and Harman (2005) later noted that this was likely because the assessors had not created the original topics themselves and were given a pre-selected set of relevant documents. Similarly, in our experiment, the judges did not make document-level relevance judgments, but were given the set of documents as chosen by the

[3] It should be noted that because the judges in our experiments did not make document-level relevance decisions, and because of the known problems associated with this (as in the TREC 2002 data sets), the validity of our data will be examined in the next section.

official TREC assessor. They also did not construct the topic query or the factual questions themselves. Therefore, we should first ensure that reasonable proportions of sentences have been selected by the judges in both the relevance and novelty steps of the task and in both settings.

The first column of Table III shows the median proportion of sentences marked as being relevant, and the median proportion of relevant sentences that were also novel, over both topics and all three judges in the topic-based setting. The proportion of relevant sentences reported on event clusters in the TREC 2003 data was 0.47, and the proportion of novel, relevant sentences was 0.61. Therefore, we conclude that the relative proportions of sentences selected in the two-stage task by our judges in the topic-based setting, do not cause concern.

In the second column, the median proportion of sentences over all 20 questions (for N4 and N33) and the three judges that were relevant and novel are shown. The median proportion of sentences in a data cluster that are relevant to a question is 1.3%, which indicates that for a given question, about 10 sentences provide an answer. There were no questions for which any of the three annotators failed to find relevant sentences. In addition, on average, one-third of the sentences that were relevant to a question were novel. Given that the judges in the test setting of our experiment are searching for sentences containing answers to specific questions, an average of 10 answers provided per question over the 25 news articles, with 3 being unique, seems quite reasonable.

	Topic-focused	Fact-focused
Relevant	0.38	0.013
Novel	0.60	0.33

Table III: Median proportion of sentences judged as being relevant/novel over all topics, questions (in the fact-focused case) and assessors.

Relevance judgments

We now turn to analyzing the agreement between annotators on the task of finding relevant sentences. Table IV shows the agreement between the three judges in the control setting. We considered the agreement with respect to the set of relevant sentences found by the official assessor, A. The figures shown are the total number of sentences in each topic, the number of relevant sentences found by A, and the proportion of sentences on which all three judges agreed. As can be seen, the three-way agreement over both topics is 15.7%. The corresponding Cohen's Kappa for this level of agreement is also 0.157, indicating that the agreement is greater than what would be expected by chance alone (Cohen, 1960). However, the level of agreement is quite clearly not very satisfactory. The reported median agreement over all event clusters on the TREC 2003 data was 0.82. However, we cannot make direct comparisons since the agreement on individual topics was not reported. In addition, the TREC tasks involved two rather than three judges.

	Total Sentences	# Relevant Sentences (Judge A)	3-way Agreement
N4	928	640	0.181
N33	708	530	0.128
All	1,636	1,170	0.157

Table IV: Topic-focused relevance judgments: proportion of judge A's relevant sentences chosen by both judge B and C.

	Total Sentences	# Relevant Sentences (Judge D)	3-way Agreement
N4	928	118	0.548
N33	708	87	0.515
All	1,636	205	0.515

Table V: Fact-focused relevance judgments: proportion of judge D's relevant sentences chosen by both judge E and F.

In considering the level of agreement on the fact-focused relevance judgments, we took the list of judge D's sentences that are relevant to a given question, and found the proportion on which the other two judges both agreed. The median, three-way agreement over the ten questions in each cluster, and the median agreement over all 20 questions are shown in Table

V. The corresponding Kappa statistic, considering the judgments over all twenty questions, is 0.53.

It is not surprising to see some variance in the level of agreement across questions. It is intuitive that questions for which there is typically only one unique answer reported in the news articles, should exhibit more agreement on relevance judgments. For example, there was perfect agreement on the question “Where was the flight's destination?" for the Egypt Air topic, and on “Where was Putin during the rescue operation?" for the Kursk topic. These questions were rather non-controversial in that only one answer was reported for each over the 25 relevant news articles. To contrast, for the question “What was the problem with the [Egypt Air] plane?" the judges did not agree on any relevant sentences. In articles about the Egypt Air crash topic, the problem or cause of the crash was never clearly reported, as it had not yet been determined though the ongoing investigation. Similarly, there was very low agreement in finding sentences relevant to the question “What caused the Kursk to sink?" since a variety of explanations were offered in the news reports. Again, in the Kursk story, a confirmed, final answer to the question was not given in any of the articles.

In summary, we confirmed our first hypothesis. The median three-way agreement on relevance judgments in the fact-focused setting was 0.515, versus 0.157 in the topic-focused setting. In addition, in the fact-focused approach, we obtained a Kappa of 0.53, which indicates a statistically significant level of agreement between the three annotators. However, as previously discussed, this is indicative of a low-to-moderate level of agreement according to most scales. In addition, we should qualify our conclusions regarding our first hypothesis in that there is a significant level of interjudge agreement when we take our set of questions as a whole. The level of agreement on relevance judgments varies across questions, and is often relatively low for questions to which no final answer was reported in the respective news articles. Therefore, it is not the case that one can necessarily expect high levels of agreement on any individual question. However, given a set of factual questions as a representation of a topic, reasonable agreement can be achieved on identifying relevant sentences over the set of questions.

Novelty judgments

Table VI describes the level of agreement between the three judges in the control setting, on the task of choosing which of the relevant sentences are also novel. Again, we used the set of novel sentences identified by judge A, as the reference set. Over both topics, 11% of the sentences chosen by A were also labeled by both B and C as being novel, corresponding to a Kappa of 0.11, since the likelihood of three judges randomly agreeing on the respective number of sentences is negligible.

To contrast, in the fact-focused setting, both judges E and F chose one-third of D's novel sentences (Kappa of 0.333). The median proportion of sentences on which all three judges agreed with respect to novelty status, are shown in Table VII for both topics and over all twenty questions. Therefore, at a first glance, it appears that the agreement on novelty status is slightly better in the fact-focused case as compared to the topic-focused setting. However, the way that we have calculated the agreement on the identification of novel sentences, by finding the proportion of sentences deemed novel by the official judge that are also considered novel by the other two judges, embeds the level of agreement that was achieved on the first step of finding relevant sentences. Since, by the definition of the novelty task, assessors choose novel sentences from their set of relevant sentences, disagreement on this first step is carried over to the second step.

In order to get an idea as to how much agreement there is on the second step of the task alone, we re-examined the data, considering only the sentences that were considered relevant to the given topic or fact, by all three judges. This filters out the disagreement on relevance judgments. Table VIII displays the number of sentences judged as being relevant by all three assessors in each setting, and the proportion of those relevant sentences found to be novel by all three assessors. As can be seen, for this conditional measure of agreement, given that the judges agreed on relevance status, their consensus on novelty status was actually better in the topic-focused case.

	Total Sentences	# Novel Sentences (Judge A)	3-way Agreement
N4	928	489	0.117
N33	708	398	0.103
All	1,636	887	0.110

Table VI: Topic-focused novelty judgments: proportion of judge A's novel sentences chosen by both judge B and C.

	Total Sentences	# Novel Sentences (Judge D)	3-way Agreement
N4	928	32	0.267
N33	708	25	0.184
All	1,636	57	0.333

Table VII: Fact-focused novelty judgments: proportion of judge D's novel sentences chosen by both judge E and F.

	# Sentences relevant	3-way Agreement
Topic-based	184	0.73
Fact-based	109	0.46

Table VIII: Number of sentences deemed relevant by all 3 judges and the proportion of these sentences judged as being novel by all 3 judges.

Therefore, we conclude that agreement is not necessarily better when judges evaluate novelty in the context of specific facts as compared to a broader topic. Agreement on relevance is built into the level of agreement on novelty assessments, given the nature of the task. Because of this, in our experiment, we observed a greater level of agreement in the fact-based setting (0.333) versus the topic-based case (0.11). Previously, researchers noted that making the sentence-level relevance judgments is likely the more difficult of the two steps in the novelty task (Allan, 2003). Our results support this claim, in that we obtained a low level of agreement in the topic-focused relevance detection task, but yet a relatively high level of agreement on the novelty judgments (0.73), for the sentences that were relevant for all three judges.

Comparing the approaches

We now turn to comparing the nature of the two approaches to novelty detection. Since we have seen that the first step in the task appears to be the more crucial, we focus on comparing the relevance judgments made under the two approaches in our experiment. In particular, we wish to investigate whether there is any evidence that the topic-based assessors adopt a fact-focused approach to finding relevant sentences. In addition, we will examine to what extent the ten questions we used for each news topics adequately represent the essence of the story.

We compared the annotations of judges A and D, the official assessors for the two approaches. D's set of relevant sentences consisted of all relevant sentences over all ten questions for each topic. In Table IX, we provide the number of relevant sentences found by each judge for each topic, as well as the number of sentences common to both judges' relevance sets. As can be seen, A recalled 89% of D's relevant sentences, despite not having used the set of questions. To contrast, D recalled only 16% of A's sentences.

	Relevant sentences (Judge A)	Relevant sentences (Judge D)	# Sentences in both sets
N4	640	118	103
N33	530	87	80
All	1,170	205	183

Table IX: Relevant sentences found by Judges A and D and the sentences deemed as relevant by both A and D.

We analyzed the 22 sentences that were labeled as relevant by Judge D but were not chosen by A. In the Egypt Air topic (N4), there were 15 such sentences. The missed sentences expressed answers to four of the ten questions, meaning that six questions were fully covered. Among the 7 sentences that were relevant to the questions surrounding the Kursk submarine topic (N33) but were not chosen by A, there were answers to three of the ten questions. Therefore, of the 20 questions used in the experiment, A actually fully covered 13 of the 20 questions, without being explicitly asked to do so.

Obviously, given the difference in the number of relevant sentences found by the two judges, a set of ten factual questions cannot fully represent the substance of a news event topic. Figure 3 shows the narrative portion of the TREC query for topic N4, which complements the title and description portions of the query previously shown in Figure 1. Clearly, it can be seen that many additional factual questions surrounding the Egypt Air crash event could be formulated from this narrative. For example, several aspects of the topic, including the reactions of family members and statements from the FAA and NTSB, were not represented in our question set.

> Details, technical and otherwise regarding the incident (e.g. number of passengers aboard, number killed, date, time, location, nationalities of victims, crew members, radio contact, radar sightings, rescue efforts and findings) are relevant. Reaction of family members and loved ones regarding the victims are relevant. Investigatory details concerning technical reasons for the crash are relevant. Analysis of recovered items associated with the incident, and the ensuing comments, opinions, findings and reports are relevant. Actions, opinions, and statements from FAA and NTSB, as well as Egyptian CAA personnel regarding the incident including warnings received prior to, and theories concerning the tragedy are relevant. Statements from Machinist Union personnel attesting to the fitness of the plane assembled by their mechanics are relevant.

Figure 3: Narrative portion of TREC topic query for N4, Egypt air crash.

In summary, the official judge in the topic-focused setting found the majority of the fact-focused judge's relevant sentences. The scope of the TREC topic queries subsumes that of our small set of questions. As seen in the narrative for the Egypt Air query in Figure 3, many of the same aspects of the story are mentioned, in addition to many other details of the story. However, we argue that it is because of the broad nature of the topic that interjudge agreement was shown to be lower in the topic-focused case, as compared to the fact-focused case. It is arguably more difficult for judges to evaluate the relevance status of sentences with respect to an entire set of facts simultaneously (i.e. a topic) rather than with respect to one fact at a time.

Limitations

Here we note the constraints of our current work. First, as the study is exploratory in nature, and given that annotating 10 facts per topic is quite a laborious procedure, we used only two event news topics in our experiments. In comparison, the TREC experiments involved a set of 50 topics of a much wider variety of subjects, and included both event and opinion stories. Since our goals were to explore the feasibility of the fact-focused approach to the task, and to compare it to the topic-focused setting, we chose to work with a specific type of news topic - emergency situations - rather than trying to cover a wider range of stories. We worked in depth using two topics, and included three judges in each of our experimental settings, in order to conduct a robust test of the interjudge agreement on the task. In addition, we focused only on developing an approach that can be used with event-type news topics, and we have not yet considered if or how our approach might apply to opinion topics.

Another factor is that in our fact-focused experiments, we used a limited set of ten questions for each news topic. As pointed out earlier, it is clear that a much larger set of questions could be developed for each topic, in order to more thoroughly represent the overall topic. However, our intent was to see how well judges agreed on finding the central facts of each news story. In other words, the questions were written with the basic “who, what, when, and where" of a story in mind. In comparison, the TREC topics are much broader, covering much more information about a topic, and many fine details of a story.

Conclusion

The sentence-level novelty detection problem is challenging for a number of reasons. The concept of novelty is inherently difficult to define and operationalize (Soboroff, 2005), in part due to the fact that it is so context sensitive (Allan, 1999). In addition, creating truth data sets for this task, which are crucial for the development and evaluation of systems, is challenged by the fact that novel sentences must first be relevant, and that relevance judgments are known to be somewhat idiosyncratic. In proposing a fact-focused approach to novelty detection, which restricts the context the judges must consider, we hoped to show that higher

levels of interjudge agreement can be obtained on both the relevance and novelty judgments. This intuition is consistent with linguistic research that has demonstrated that constraining context produces more similar word associations among subjects (e.g. (Aitchison, 2003)). In the case of our problem, considering relevance at the factual level rather than at the broader, topical level might mean that subjects have similar interpretations of the meaning of sentences, thus resulting in more congruent relevance judgments. Finally, we also reasoned that considering relevance and novelty in the context of a single fact would be a clearer procedure than considering multiple facets of a story, as in the case of topic-focused judgments. This is because we were able to frame the fact-focused problem as a search for answers to factual questions surrounding a given topic, which could be tackled one at a time by the subjects.

In our experiments, there was indeed a higher level of agreement on the relevance judgments in the fact-focused case as compared to the topic-focused setting, when we considered the average level of agreement over the entire set of questions. However, as noted in our analysis, the extent of agreement does depend on the type of question asked. In particular, finding answers to more open-ended questions, such as those concerning the cause of an accident under investigation, is more difficult as compared to questions that have very precise answers, or to which only one unique answer is reported in the set of documents. Therefore, how much the ground truth surrounding a story fluctuates, and how definite news reports of a situation are, clearly affect the extent to which users agree on which sentences are relevant.

Our results on the novelty portion of the task were interesting and unexpected. When we calculated the agreement on novelty judgments as done in the TREC evaluation, in which the agreement on relevance is inherently embedded, the fact-focused case again had a higher level of agreement as compared to the topic-focused approach. However, when we considered only the sentences on which all three judges agreed as to relevance status, the agreement was higher in the topic-focused case. This seems to confirm the view that the task of identifying the sentences that are relevant to a topic is more difficult than that of determining which topic-relevant sentences are also novel.

In summary, while constraining the context in which relevance judgments are made appears to improve interjudge agreement, this does not appear to be true with respect to novelty judgments. In future work, it may be fruitful to examine the two facets of the problem separately, and to consider if different approaches to the sentence-level relevance and novelty detection problems have useful applications within current IR systems. As an example, one can imagine an online text summarization system designed to support users in following breaking news stories, such as the ones examined in the current study. Such a system might collect related news articles about a given topic as they are published over time and then provide summaries designed to help users understand the key elements of the story as it evolves. Here, different approaches could be implemented. The system could provide question-focused summaries in response to users' input factual questions, over time as demanded by the user. Thus, such a system would focus on retrieving fact-relevant sentences, while allowing the user to decide for herself as to the novelty of the information presented. Alternatively, topic-focused -summaries (Allan, 2001) could be produced, in an effort to highlight what information has newly become available since the user's last interaction with the system. An interesting direction for further work could compare these two approaches in terms of how well they support users in understanding complex, time-sensitive information as presented in a set of multiple text documents written over time.

Acknowledgements

This work was partially supported by the U.S. National Science Foundation under the following grant: 0329043 ``Probabilistic and Link-based Methods for Exploiting Very Large Textual Repositories'' administered through the IDM program. All opinions, findings, conclusions, and recommendations in this paper are made by the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the members of the CLAIR research group at the University of Michigan and the anonymous Journal of Documentation reviewers for their feedback and comments on this work.

References

Aitchison, J. Words in the Mind: An Introduction to the Mental Lexicon. Blackwell Publishing Ltd., 2003.

Allan, J., Carbonell, J., Doddington, G., Yamron, J and Yang, Y. (1998), "Topic detection and tracking pilot study final report", *Proceedings of the Defense Advances Research Projects Agency (DARPA) Broadcast News Transcription and Understanding Workshop,* pp. 194-218.

Allan, J., Carterette, B., and Lewis, J. (2005), "When will information retrieval be "good enough"?" *Proceedings of the 28th Annual Association for Computing Machinery Conference on Research and Development in Information Retrieval,* Association for Computing Machinery, Salvador, Brazil.

Allan, J., Gupta, R., and Khandelwal, V. (2001), "Topic models for summarizing novelty", *Proceedings of the Workshop on Language Modeling and Information Retreival,* pp. 66-71, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Allan, J., Jin, H., Rajman, M, Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R. and Caputo, D. (1999), "Topic-based novelty detection 1999 summer workshop at Center for Language and Speech Processing final report," Johns Hopkins University, Baltimore, Maryland.

Allan, J., Wade, C. and Bolivar, A. (2003), "Retrieval and novelty detection at the sentence level," *Proceedings of the 26th Annual Association for Computing Machinery Conference on Research and Development in Information Retrieval,* Association for Computing Machinery, Toronto, Canada.

Amigo, E., Gonzalo, J., Peinado, V., Penas, A. and Verdejo, F. (2004), "An empirical study of information synthesis task", *Proceedings of the 42nd Meeting of the Association for Computational Linguistics,* pp. 207-214, Association for Computational Linguistics, Barcelona, Spain.

Carletta, J. (1996), "Assessing agreement on classification tasks: the Kappa statistic", *Computational Linguistics,* Vol. 22(2), pp. 249-254.

Cohen, J. A. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement,* Vol. 20, pp. 37-46.

Cuadra, C. A. and Katter, R. V. (1967), "Opening the black box of relevance", *Journal of Documentation,* Vol. 23(4), pp. 291-303.

Di Eugenio, B. (2000), "On the usage of Kappa to evaluate agreement on coding tasks," *Proceedings of the Second International Conference on Language Resources and Evaluation,* pp. 441-446, Athens, Greece.

Ficus, J. G. and Doddington, G. R. (2002), "Topic detection and tracking evaluation overview", *Topic Detection and Tracking: Event-based Information Organization,* pp. 17-32, Kluwer Academic Publishers.

Froehlich, T. J. (1994), "Relevancy reconsidered – towards an agenda for the 21st century: introduction to the special topic issue", *Journal of the American Society for Information Science,* Vol. 45(3), pp. 124-133.

Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J. (1999), "Summarizing text documents: sentence selection and evaluation metrics", *Proceedings of the 22nd Annual Association for Computing Machinery Conference on Research and Development in Information Retrieval,* Association for Computing Machinery, Berkeley, California.

Harman, D. (2002), “Overview of the TREC 2002 novelty track”, *Proceedings of the 11th Text Retrieval Conference,* National Institute of Standards and Technology, Gathersburg, Maryland.

Harter, S. P. (1996), “Variations in relevance assessments and the measurement of retrieval effectiveness”, *Journal of the American Society for Information Science,* Vol. 47(1), pp. 37-49.

Krippendorff, K. (1980), *Content Analysis: An Introduction to its Methodology*, Sage Publications, Beverly Hills, California.

Landis, J. R. and Koch, G. G. (1977), “The measurement of observer agreement for categorical data”, *Biometrics,* Vol. 33, pp. 159-174.

Marcu, D. and Echihabi, A. (2002), “An unsupervised approach to recognizing discourse relations”, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Philadelphia, Pennsylvania.

Mizzaro, S. (1997), “Relevance: the whole history”, *Journal of the American Society for Information Science,* Vol. 48(9), pp. 810-832.

Nenkova, A. and Passonneau, R. (2004), “Evaluating content selection in summarization: the pyramid method”, *Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technology Conference,* North American Association for Computational Linguistics, Boston, Massachusetts.

Schamber, L. (1994), “Relevance and information behavior”, *Annual Review of Information Science and Technology,* Vol. 29, pp. 3-48.

Schiffman, B. (2005), *Learning to Identify New Information,* PhD Dissertation, Department of Computer Science, Columbia University, New York, New York.

Siegel, S. and Castellan, N. J. (1998), *Nonparametric Statistics for the Behavioral Sciences,* McGraw Hill Publishers.

Soboroff, I. and Harman, D. (2003), “Overview of the TREC 2003 novelty track”, *Proceedings of the 12th Text Retrieval Conference,* National Institute for Standards and Technology, Gaithersburg, Maryland.

Soboroff, I. and Harman, D. (2005), “Novelty detection: the TREC experience”, *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Vancouver, Canada.

Teufel, S. and Moens, M. (2002), “Articles summarizing scientific articles: experiments with relevance and rhetorical status”, *Computational Linguistics,* Vol. 28(4).

Vakkari, P. and Sormunen, E. (2004), “The influence of relevance levels on the effectiveness of interactive information retrieval”, *Journal of the American Society for Information Science and Technology*, Vol. 55(11), pp., 963-969.

van Halteren, H and Teufel, S. (2003), “Examining the consensus between human summaries: initial experiments with factoid analysis”, Proceedings *of Human Language Technology and North American Association for Computational Linguistics Workshop on Text Summarization.* Association for Computational Linguistics, Edmonton, Canada.

Voorhees, E. M. (2000), “Variations in relevance judgments and the measurement of retrieval effectiveness”, *Information Processing and Management,* Vol. 36(5), pp. 697-716.

Voorhees, E. and Tice, D. (2000), “The TREC-8 question answering track evaluation”, *Proceedings of the 8th Text Retrieval Conference,* National Institute for Standards and Technology, Gaithersburg, Maryland.

Zhang, C., Callan, J. and Minka, T. (2002), “Novelty and redundancy detection in adaptive filtering”, *Proceedings of the 25th Annual Association for Computing Machinery Conference on Research and Development in Information Retrieval,* Association for Computing Machinery, Tampere, Finland.